Medical Image Content-Based Queries using the Grid

J. Montagnat[1], H. Duque[12], J.M. Pierson[2], V. Breton[3], L. Brunie[2], and I. E. Magnin[1]

1. CREATIS, CNRS UMR 5515, INSA, 69621 Villeurbanne, France
   http://www.creatis.insa-lyon.fr/
2. LIRIS, CNRS FRE 2672, INSA, 69621 Villeurbanne, France
   http://liris.insa-lyon.fr/
3. LPC, CNRS-IN2P3, Campus des Cézeaux, 63177 Aubière Cedex, France
   http://clrwww.in2p3.fr/

Corresponding author:
Johan Montagnat
CREATIS, INSA, Bât. B. Pascal
20 av. A. Einstein
69621 Villeurbanne Cedex
France

email: johan@creatis.insa-lyon.fr
phone: +33 492 94 27 20
fax: +33 492 94 28 98

Running title: Medical Image Content-Based Queries
Keywords: Medical imaging, grid computing, medical database, content-based query

Medical Image Content-Based Queries using the Grid

J. Montagnat[1], H. Duque[12], J.M. Pierson[2], V. Breton[3], L. Brunie[2], I. E. Magnin[1]
[1] CREATIS, CNRS UMR 5515, INSA, 69621 Villeurbanne, France, http://www.creatis.insa-lyon.fr/
[2] LIRIS, CNRS FRE 2672, INSA, 69621 Villeurbanne, France, http://liris.insa-lyon.fr/
[3] LPC, CNRS-IN2P3, Campus des Cézeaux, 63177 Aubière Cedex, France, http://clrwww.in2p3.fr/

Abstract

Computation and data grids have encountered a large success among the scientific computing community in the past few years. The medical imaging community is increasingly aware of the potential benefit of these technologies in facing today medical image analysis challenges. In this paper, we report on a first experiment in deploying a medical application on a large scale grid testbed. Our pilot application is a hybrid metadata and image content-based query system that manipulates a large data set and for which image analysis computation can be easily parallelized on several grid nodes. We analyze the performances of this algorithm and the benefit brought by the grid. We further discuss possible improvements and future trends in porting medical applications to grid infrastructures.

1 Introduction

Digital medical images represent tremendous amounts of data for which automatic processing and intelligent indexing is increasingly needed [10, 1]. The annual production of a radiology department in an industrialized country hospital is in the order of 10 Terabytes. The total data produced in European countries is therefore in the order of Petabytes per year and the total medical data of Europe or the USA can be estimated to thousands of Petabytes. Although often coming from digital sources, a large majority of these data is usually not archived in the long term.

Grids [4, 5] are a promising tool to build medical databases and to face health-related challenges involving computations over large datasets such as epidemiology, image content-based retrieval, or drug assessment. Indeed, grids offer an infrastructure for:

- sharing data and building virtual databases distributed over several medical sites;
- and sharing processing power, allowing for very large scale study involving massive data processing.

Grids have encountered a large success among scientific computing communities. However, the medical fields related requirements have not been so well analyzed so far. Today, grids provide a low level hardware and software architecture but hardly address the high level user interfaces and services mandatory for deploying medical applications over a wide scale such as security and distribution of medical data.

In this paper, we focus on an image indexing and querying application over large databases [8] that was experimented on the DataGrid European IST project testbed. The application allows to:

- Register a set of medical images and their associated metadata on a grid data manager.

- Search these data by making hybrid queries involving both metadata and image contents.

The image content-based queries involved classical image similarity measures (section 2). A prototype deployed on a limited infrastructure is first demonstrated (section 3). An architecture allowing a more integrated implementation and taking into account more realistic constraints is then discussed (section 4).

2 Image content-based queries

The overall application is depicted in figure 1. This system allows to (i) register medical images coming from medical imagers and their medical data and (ii) to make queries such as '*search for Magnetic Resonance Images of the heart similar to the one of Mr Foo Bar acquired yesterday in my hospital*'. The image can then be visualized on the physician screen and compared to the most similar cases available from the database. Figure 1 decomposes this application in 7 steps:

1. Storage of medical images coming from an imager on large capacity disk servers.
2. Indexing of images according to patient-specific and image-related metadata such as the patient name, the acquisition type, etc.
3. Retrieval of images for visualization on the physician screen. Queries can involve both metadata and image contents. Metadata are stored in a standard database queried using the SQL language.
4. Query over image content involves triggering automatic processings.
5. The image database is transported to computation nodes.
6. Similarity measures are used to discover images that show similar content to a given sample.
7. The most similar images to the case of study can then be visualized on the physician screen.

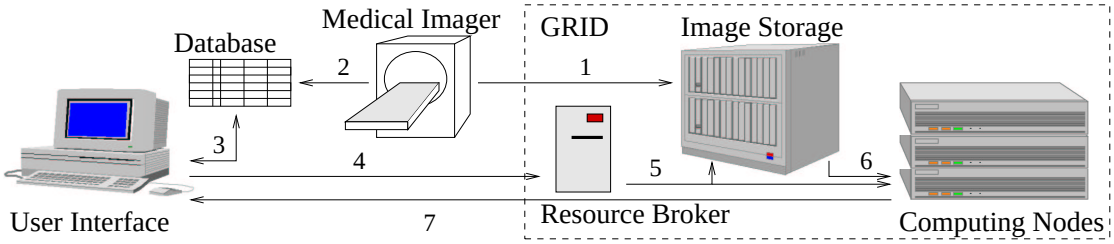

Figure 1. Application synopsis

Given an image of interest (the source image), the hybrid query system performs first a pre-selection of possible images to compare (the target images) by querying the metadata and then an image content-based analysis on the targets to search for most similar images. The pre-selection is based on a simple SQL database query. Only images with comparable attributes (*e.g.* same region of interest as the source image, same modality, same orientation, etc) are considered as possible targets. The resulting data set can be quite large though (up to few hundred images in our experiments, may be much more in clinical practice). Therefore, the computing power of a grid is used to perform the content-based analysis step. Each image in the target set is compared to the source image through a similarity measure [6, 7]. The resulting score is used to rank the target images from most to less similar.

2.1 Image analysis

Several similarity measurements have been implemented in our application. Let I represent the source image and J represent one target image with the same support (both I and J have n voxels). Let i denote the gray level intensity of voxels in image I and j denote the the gray level intensity of voxels in image J. n_i (resp. n_j) is the number of voxels with intensity i (resp. j) in image I (resp. J). n_{ij} is the number of voxels having simultaneously intensity i in image I and j in image J. $p_i = \frac{n_i}{n}$ and $p_j = \frac{n_j}{n}$ are associated probabilities. $p_{ij} = \frac{n_{ij}}{n}$ is the joint probability of a voxel at a given location to have intensity i in I and j in J. The mean and variance of intensities in image I can be computed as: $m_I = \sum_i i p_i$ and $\sigma_I^2 = \sum_i (i - m_I)^2 p_i$. From these statistical measurements, one can compute several similarity measures:

- **The simple differences:**

$$D_1(I,J) = \sum_i \sum_j p_{ij}|i-j| \text{ and } D_2(I,J) = \sum_i \sum_j p_{ij}(i-j)^2 \tag{1}$$

  are simple measurements for mono-modal image comparisons. They are sensitive to signal noise and inhomogeneities so their principal interest is their simplicity.

- **The coefficient of correlation:**

$$\rho^2(I,J) = \frac{\mathrm{Cov}^2(I,J)}{\mathrm{Var}(I)\mathrm{Var}(J)} = \sum_i \sum_j \frac{(i-m_I)(j-m_J)}{\sqrt{\sigma_I}\sqrt{\sigma_J}} \tag{2}$$

  is a normalized measurement taking into account an affine transformation between I and J intensities. It has been extensively used in the literature.

- **The Wood's criterion:**

$$\mathrm{W}(I,J) = \sum_j \frac{\sigma_{I|j}}{m_{I|j}} p_j \text{ with } \left\{ \begin{array}{l} m_{I|j} = \frac{1}{p_i} \sum_j j p_{ij} \\ \sigma^2_{I|j} = \frac{1}{p_i} \sum_j (j - m_{J|i})^2 p_{ij} \end{array} \right. \tag{3}$$

  was introduced to register MRI on PET images. Given the set of voxels with intensity i in the source image, the Wood's criterion measures the variation of intensities of corresponding voxels in the target image.

- **The correlation ratio:**

$$\mu^2(I,J) = 1 - \frac{1}{\sigma_I^2} \sum_j p_j \sigma^2_{I|j} \tag{4}$$

  is used for multi-modal registration and makes the hypothesis that a functional relation exists between the source and the target image intensities.

- **The mutual information**, or entropy:

$$H(I,J) = -\sum_i \sum_j p_{ij} \frac{p_{ij}}{p_j} \tag{5}$$

  is the most general similarity measure. It measures the entropy of the joint gray levels distribution without any assumption on an existing relation between source and target image intensities.

2.2 Computations complexity

All above mentioned measures computation require to build the joint histogram of the two images and to compute the statistics. The joint histogram in an $m \times m$ sparse matrix where m is the number of possible gray levels (typically, $m = 2^{12}$ or $m = 2^{16}$ in medical images). The statistics p_i, p_j, p_{ij}, m_I, σ_I, *etc*, can then be computed.

Assembling a plain squared matrix of size m^2 with $m = 2^{16}$ is out of range of today workstation's memory capacity. Therefore:

- either the image gray levels must be undersampled in the range $[2^8, 2^{12}]$, resulting in a less precise similarity measurement,
- or a sparse matrix representation must be used which as an impact on the assembling time and the computation cost of statistics.

The former solution will be preferred for computation time while the later will be preferred for accuracy.

The computation time is also directly proportional to the image size as the joint histogram assembling algorithm is going through all image voxels in sequence. In our experiments, we used small size 2D slices of the thorax (256×256, *i.e.* $n = 65536$) and medium size 3D brain MR images ($181 \times 217 \times 181$, *i.e.* $n = 7109137$). The measured computation time of similarity measures on a 1 Ghz Pentium III processor are reported in figure 2. From this study, we will retain that the total computation time for a pair of images is a fraction of a second ($t_{e1} = 0.03$s) in the case of 2D images, about 15 seconds ($t_{e2} = 15$s) in the case of 3D images with undersampled gray levels and about 12 minutes ($t_{e3} = 720$s) with full scale 3D images.

Images	Histogram	D_1, D_2	ρ	W	μ	H
$m = 2^8$, $n = 65536$	0.030	0.020	0.033	0.031	0.030	0.040
$m = 2^{12}$, $n = 7109137$	7.90	6.56	8.28	7.97	8.15	8.03
$m = 2^{16}$, $n = 7109137$	693	4.41	7.32	9.71	9.65	11.27

Figure 2. Joint histogram and similarity measures computation time in seconds

3 Prototype

A prototype has been developed and deployed on the European DataGrid middleware to test the image content-base query application. The DataGrid application testbed offers about 1000 of CPUs spread over ten sites in Europe and a total storage capacity of a few Terabytes.

Images have first been pre-registered into the DataGrid data manager and stored on grid storage disks. To each registered image is associated a unique grid-wide identifier called LFN (Logical File Name). The associated metadata (patient-related information, hospital-related information, and image-related information) for all registered images are simultaneously recorded in an SQL data base. The LFN column in the database makes the association between the image and the metadata.

Through a graphical interface, the physician can query the database to select a source image and download it on its local machine for visualization. He can then search for similar images in the database. Figure 3 shows, from left to right, the user interface, a thorax source image, and matching images with high similarity ($\rho = 0.9$), medium ($\rho = 0.1$) and low similarity ($\rho = 0.001$) scores.

One job is started for each pair of images to compare. The grid middleware is receiving job descriptions (executable and input data), processes the jobs and returns the result (a simple similarity score). Grids clearly show a potential for such an application as all similarity measurements are independent and can be processed in parallel on different grid nodes. In clinical

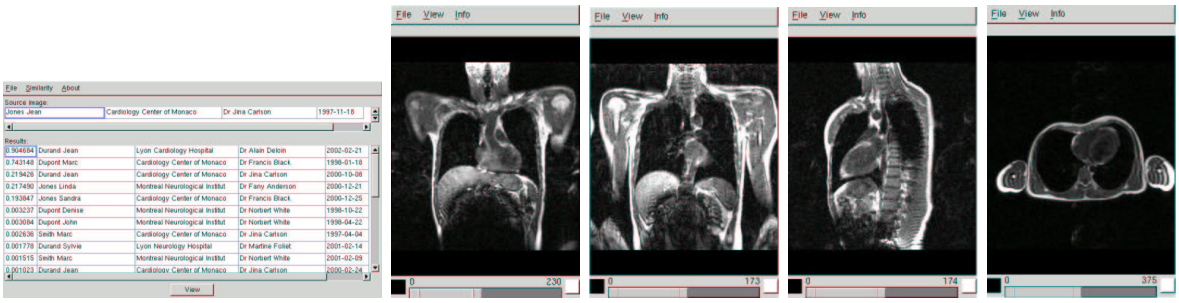

Figure 3. from left to right: application user interface, source image and matching images with highest to lowest score.

practice, an almost immediate response time would be necessary for the system to be used. However the current situation is not very satisfying as:

- Job submission requires a rather costly resource match-making computation. Currently it takes about 30 seconds to find matching resources ($t_m = 30$s).
- The system then need to transfer the job executable on the target node, wait for the job to be started from the local queue, and, once processing is done, to retrieve the job output. This introduces an additional latency for each job in the order of 60 seconds ($t_l = 60$s).
- Some current limitations of the system prevent the application to submit too many jobs at the same time to the grid resource broker. Currently, no more than 40 jobs can be executed in parallel.

Therefore, grid jobs are penalized by an incompressible additional latency that is not fully compensated by the job parallelism. Quantitatively, the sequential execution time when comparing the source image to N target images is $t_{\text{seq}} = Nt_e$ (with t_e equal to t_{e1}, t_{e2}, or t_{e3} depending on the image size and the algorithm parameters) while the grid execution time can be approximated by $t_{\text{grid}} = (t_m + t_l + t_e)\frac{N}{40} = (90 + t_e)\frac{N}{40}$. The theoretical acceleration is:

$$a = \frac{t_{\text{seq}}}{t_{\text{grid}}} = \frac{40t_e}{90 + t_e} \tag{6}$$

This function is asymptotic with a theoretical maximum of 40. The acceleration becomes positive ($a > 1$) for $t_e > 2.3$ seconds approximatively. Therefore, the gridification is valuable only for large enough execution times, independently of the number of images to match. For the three typical execution times $t_{e[1..3]}$ given above, the acceleration and the total computation time for various values of N are given in the following table:

time	acceleration	$N = 50$	$N = 100$	$N = 200$	$N = 500$
$t_{e1} = 0.03$	0.013	112	225	450	1125
$t_{e2} = 15$	5.714	131	262	535	1312
$t_{e3} = 720$	35.556	1012	2025	4050	10125

Given that in clinical practice, a computation time of 5 minutes is already a quite long time, the application is of limited interest on the current system. However, middleware improvements, and in particular breaking the limitation to 40 jobs submission would drastically change these figures.

4 Discussion and future work

We have deployed a pilot medical application on a large scale computation grid. First results reported in this paper show the potential of grids for such an application and the current limitation of the system in terms of efficiency. Concerning medical image processing, other key points were not addressed in this first prototype such as:

- **Security**. Data integrity and respect of privacy is a key issues in most applications manipulating medical images. Security is a broad concept that covers authentication of individuals and authorization checking, security of data stored and transferred over the network, logging, etc.
- **Efficient data transfers**. Latency in data transfers is a key component of the total computation time in an application such as the one reported in this paper. On a grid, data are supposed to be distributed over different sites and replicated for efficiency and robustness [9]. In out experiments, data were pre-replicated on a couple of sites and all jobs were executed on theses sites with LAN access to the input data. On a full running grid, one could imagine to get data from remote sites and use different network quality of services to optimize data transfers. For instance, the source image could sent to different sites using network broadcasting since it will be needed by all jobs.
- **Data distribution**. On a full scale system, the images should be distributed over different sites. One could imagine that each image producing center (hospital) is storing locally its images.
- **Interface with medical imagers**. In our experiment, we registered each input image on grid disks. In a fully working system, one could expect that medical imagers are automatically registering images on the grid data manager as they become available.

To deal with these different issues, we are working on a Distributed Medical Data Manager (DM^2) system [3]. The DM^2 is depicted in figure 4. It provides an interface between the medical image servers (DICOM standard [2]) and the grid middleware. It also adds capabilities to the grid data management system such as medical metadata management, data encryption and fine level authorization. Its secondary role is to provide a distributed system able to interconnect different sites, to make cross-site queries, and to retrieve medical images from different sites.

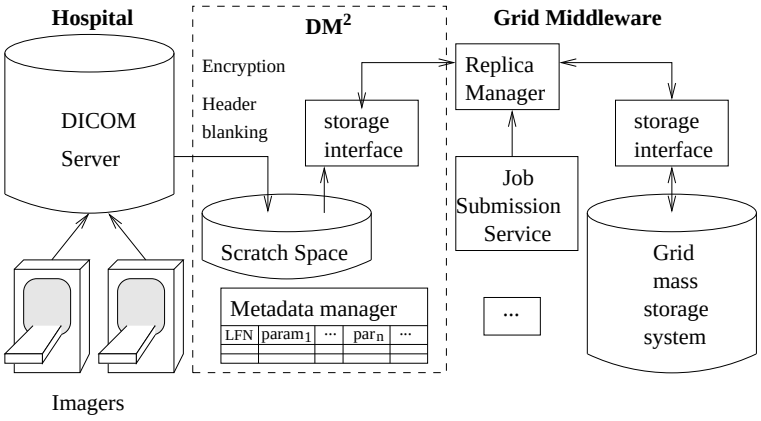

Figure 4. DM^2: A Distributed Medical Data Manager

For a clean design of this complex distributed multi-processes software component, we have drawn a layered architecture for the DM^2. On the lowest level (layer 0), a message passing kernel is allowing efficient communication between processes. On layer 1, a transactional system is ensuring coherent transactions between the different services and interfaces to external

components. Layer 2 is a distribution layer. On top of that comes an application layer with a programmable interface allowing to develop applications that take advantage of the DM^2 services and a user interface layer with high level access interface.

5 Conclusion

There is a strong potential in grids for facing medical image analysis challenges. In this paper, we reported on a pilot application that we deployed on a large scale grid testbed and we depicted some future developments. Many more applications involving large data sets (*e.g.* epidemiology studies) or computing power (*e.g.* surgery simulation) could benefit from grid technologies. However, many aspects relevant to the medical applications such as security, data management or real/limited time situations need to be further investigated. Basic middleware services are available for experiments today but there is still a large space for implementing new higher level services applications oriented.

Acknowledgments

This work is partly supported by the European DataGrid IST project, the French ministry ACI-GRID project, and the ECOS Nord Committee (action C03S02).

References

[1] R. Acharya, R. Wasserman, J. Sevens, and C. Hinojosa. Biomedical Imaging Modalities: a Tutorial. *Computerized Medical Imaging and Graphics*, 19(1):3–25, 1995.

[2] DICOM: Digital Imaging and COmmunications in Medicine. http://medical.nema.org/.

[3] H. Duque, J. Montagnat, J.M. Pierson, L. Brunie, and I.E. Magnin. DM2: A Distributed Medical Data Manager for Grids. In *Biogrid'03, proceedings of the IEEE CCGrid03*, Tokyo, Japan, May 2003.

[4] I. Foster, C. Kesselman, and S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications*, 15(3), 2001.

[5] J. Montagnat, V. Breton, and I.E. Magnin. Using grid technologies to face medical image analysis challenges. In *Biogrid'03, proceedings of the IEEE CCGrid03*, Tokyo, Japan, May 2003.

[6] G.P. Penney, J. Weese, J.A. Little, P. Desmedt, D.LG. Hill, and D.J. Hawkes. A Comparison of Similarity Measures for Use in 2D-3D Medical Image Registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'98)*, volume 1496 of *LNCS*, pages 1153–1161, Cambridge, USA, October 1998. Springer.

[7] A. Roche, G. Malandain, X. Pennec, and N. Ayache. The Correlation Ratio as a New Similarity Measure for Multimodal Image Registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'98)*, volume 1496 of *LNCS*, pages 1115–1124, Cambridge, USA, October 1998. Springer.

[8] D. Sarrut and S. Miguet. ARAMIS: A Remote Access Medical Imaging System. In *International Symposium on Computing in Object-Oriented Parallel Environments*, San Francisco, USA, December 1999.

[9] H. Stockinger, A. Samar, B. Allcock, I. Foster, K. Holtman, and B. Tierney. File and object replication in data grids. In *10th IEEE Symposium on High Performance and Distributed Computing (HPDC2001)*, August 2001.

[10] M. Thomson, W. Johnson, and Goujun G. et al. Distributed healthcare imaging information systems. In *PACS Design and Eval.: Eng. and Clinical Issues, SPIE Med. Imaging*, volume 3035, 1997.